Atten Percept Psychophys (2017) 79:2055–2063
DOI 10.3758/s13414-017-1355-0

Automatic auditory disambiguation of visual awareness

John Plass[1] · **Emmanuel Guzman-Martinez**[1] · **Laura Ortega**[1] · **Satoru Suzuki**[1,2] · **Marcia Grabowecky**[1,2]

Published online: 20 June 2017

Abstract Multisensory integration can play a critical role in producing unified and reliable perceptual experience. When sensory information in one modality is degraded or ambiguous, information from other senses can crossmodally resolve perceptual ambiguities. Prior research suggests that auditory information can disambiguate the contents of visual awareness by facilitating perception of intermodally consistent stimuli. However, it is unclear whether these effects are truly due to crossmodal facilitation or are mediated by voluntary selective attention to audiovisually congruent stimuli. Here, we demonstrate that sounds can bias competition in binocular rivalry toward audiovisually congruent percepts, even when participants have no recognition of the congruency. When speech sounds were presented in synchrony with speech-like deformations of rivalling ellipses, ellipses with crossmodally congruent deformations were perceptually dominant over those with incongruent deformations. This effect was observed in participants who could not identify the crossmodal congruency in an open-ended interview (Experiment 1) or detect it in a simple 2AFC task (Experiment 2), suggesting that the effect was not due to voluntary selective attention or response bias. These results suggest that sound can automatically disambiguate the contents of visual awareness by facilitating perception of audiovisually congruent stimuli.

Keywords Visual awareness · Multisensory processing · Binocular vision: Rivalry/Bistable perception

Despite the inherent ambiguity of sensory information, perceptual experience is typically characterized by a remarkable unity and self-consistency. One critical contributor to this perceptual reliability is information integration between the senses (Ernst & Bülthoff, 2004). Multisensory integration is thought to resolve perceptual ambiguities by facilitating awareness of percepts that are consistent across senses (Klink, van Wezel, & van Ee, 2012). As a result, in situations of high ambiguity, input from one sensory domain can crossmodally influence the contents of awareness in another modality (Schwartz, Grimault, Hupé, Moore, & Pressnitzer, 2012).

One tool frequently used to study crossmodal influences on visual awareness is binocular rivalry. In binocular rivalry, incompatible images are presented to corresponding locations in each eye, resulting in spontaneous alternation between the images in visual awareness (Blake & Logothetis, 2002; Wheatstone, 1838). While these alternations generally occur stochastically (e.g., Blake, Fox, & McIntyre, 1971; Kim, Grabowecky, & Suzuki, 2006), multisensory information can increase the overall perceptual dominance of crossmodally congruent percepts (e.g., Lunghi, Binda, & Morrone, 2010; Zhou, Jiang, He, & Chen, 2010).

Several studies using binocular rivalry suggest that auditory stimuli can bias visual awareness toward audiovisually congruent percepts (Chen, Yeh, & Spence, 2011; Conrad, Bartels, Kleiner, & Noppeney, 2010; Guzman-Martinez, Ortega, Grabowecky, Mossbridge, & Suzuki, 2012; Kang & Blake, 2005; Parker & Alais, 2006). However, further results suggest that these effects may require voluntary selective attention to congruent visual stimuli (van Ee, van Boxtel, Parker, & Alais, 2009) and, therefore, may not reflect disambiguation through

✉ John Plass
johncplass@u.northwestern.edu

[1] Department of Psychology, Northwestern University, Swift Hall 102, 2029 Sheridan Road, Evanston, IL 60208, USA

[2] Interdepartmental Neuroscience Program, Northwestern University, Swift Hall 102, 2029 Sheridan Road, Evanston, IL 60208, USA

crossmodal facilitation. Rather, because voluntary attention can prolong rivalry dominance (Chong, Tadin, & Blake, 2005; Lack, 1974), these effects may reflect voluntary shifts in visual attention toward congruent stimuli. If this were the case, previous results would not reflect multisensory interactions in perceptual processing but simple auditory cuing of an attentional target. Audiovisual correspondences were easy to detect in most prior studies (e.g., synchronous flicker and amplitude modulation), leaving open the possibility that recognition of audiovisual congruency led participants to disproportionally direct selective attention toward congruent stimuli (Deroy, Chen, & Spence, 2014; Klink et al., 2012). Further, such recognition could have biased participant towards reporting mixed percepts as target stimuli due to demand characteristics.

However, contrary to these explanations, research from our laboratory suggests that visual awareness favors audiovisually congruent percepts, even when the congruency is not apparent to participants. Guzman-Martinez and colleagues (2012) found that Gabor patches dominated in binocular rivalry when their spatial frequencies corresponded to the amplitude modulation rate of simultaneously presented sounds. In postexperimental interviews, none of the participants reported recognizing any relationship between the Gabor patches and the sounds, suggesting that crossmodal disambiguation, and not attentional or response biases, drove this effect. Here, we sought to confirm this interpretation by demonstrating a similar effect using novel audiovisual speech stimuli and a more stringent behavioral test of congruency detection.

Audiovisual speech stimuli are some of the most likely multisensory stimuli to be strongly integrated in perception (Navarra, Yeung, Werker, & Soto-Faraco, 2012). Infants as young as 2 months of age can detect matches in phonetic content between auditory and lip-read speech (Baier, Idsardi, & Lidz, 2007; Patterson & Werker, 2003). In adults, the perceptual system is particularly forgiving of temporal asynchronies between auditory and visual speech, suggesting that it may be especially likely to assume that audiovisual speech stimuli in particular originate from the same source (Vatakis, Ghazanfar, & Spence, 2008; Vatakis & Spence, 2007, 2008).

Previously, Sweeny and colleagues (Sweeny, Guzman-Martinez, Ortega, Grabowecky, & Suzuki, 2012) found that speech sounds crossmodally influenced visual perception of ellipses without participants' knowledge of an association between the stimuli. Taking advantage of this typically unnoticed association, we created animated ellipses that mimicked the dynamics of speaking mouths and presented them under rivalry conditions. The animation presented to one eye changed shape in a manner consistent with simultaneously presented auditory speech, whereas the animation presented to the other eye moved inconsistently. After recording participants' perceived dominance throughout binocular rivalry, we verified that they were naïve to the audiovisual congruence (with an open-ended interview in Experiment 1 and using a two-alternative-forced-choice, 2AFC, task in Experiment 2), ensuring that any observed effects could not be driven by voluntary selective attention or response bias toward the audiovisually congruent animation.

Even when participants did not recognize the relationship between the auditory and visual stimuli, the audiovisually congruent stimulus dominated in binocular rivalry. These results suggest that auditory input can automatically disambiguate the contents of visual awareness by facilitating perception of audiovisually congruent stimuli.

Experiment 1

Based on research suggesting that speech sounds are integrated with ellipses without participants' recognition of their relationship (Sweeny et al., 2012), we created a binocular rivalry display consisting of ellipses that changed shape either consistently or inconsistently with concurrent speech sounds. We expected that phonemes produced by horizontal extensions of the lips would be integrated with horizontally extending ellipses, whereas phonemes produced by vertical extensions of the lips would be similarly integrated with vertically extending ellipses. In our binocular rivalry display, the ellipse animation presented to one eye repeatedly expanded horizontally and vertically in a manner consistent with concurrent /wi:/ (IPA transcription; as in English: "we") and /wo / (as in English: "woe") speech sounds. The ellipse animation presented to the other eye, however, moved incongruently, expanding vertically during the /wi:/ sound and horizontally during /wo / sounds (see Fig. 1). We recorded dominance durations for the rivaling animations in order to determine the effect of audiovisual congruency on visual awareness.

Method

Participants Twelve undergraduate students at Northwestern University gave informed consent to participate for partial course credit. They all had normal or corrected-to-normal visual acuity, normal color vision, and normal hearing. Each was tested individually in a dimly lit room.

Apparatus Visual stimuli were presented on a 21-in. color CRT monitor (85 Hz, 1400 × 1050 resolution) at a viewing distance of 110 cm. A stereoscope with four front-surface mirrors, a central divider, and an integrated headrest was used to present different animations to each eye. Sounds were presented via a pair of JBL speakers (10–25000 Hz frequency response) placed symmetrically just beneath the monitor.

Stimuli Visual stimuli consisted of two dichoptically presented animations of ellipses that continuously elongated

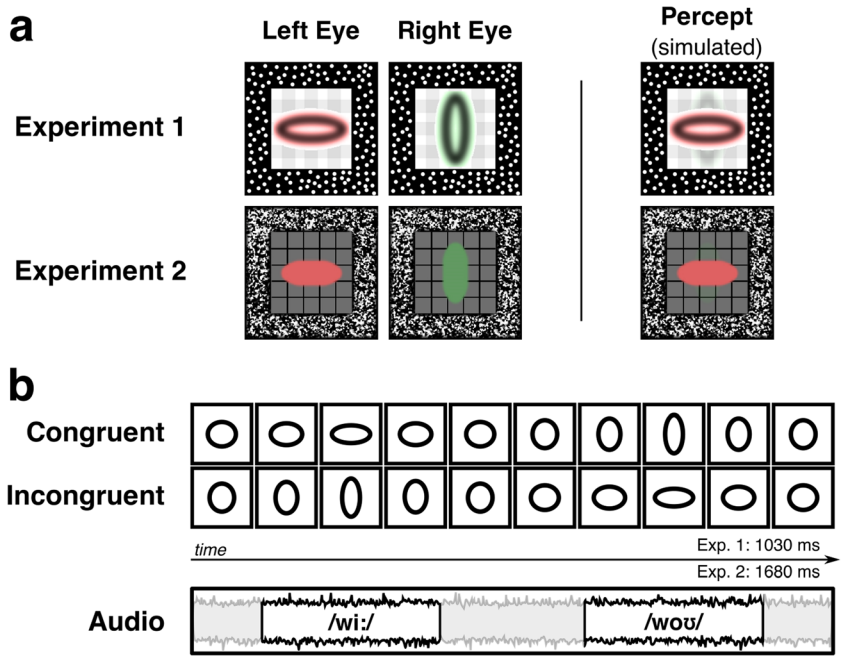

Fig. 1 Stimulus overview. **a** Left: Sample frames from the binocular animations presented in Experiment 1 (*top*) and Experiment 2 (*bottom*). Ellipses dynamically alternated between horizontal and vertical expansion, while the binocularly presented frames and gridded background remained static to facilitate stable binocular fusion. Right: A potential percept produced by the sample frames. During actual viewing, stimulus dominance and the degree of suppression varied stochastically. **b** Schematic representation of the audiovisual stimuli in Experiment 1 and the synchronous condition in Experiment 2 (timing approximate). This sequence was repeated several times during each trial. As in natural speech, the congruent ellipse expanded horizontally in synchrony with the /wi:/ sound and vertically with the /wo / sound. The incongruent ellipse was rotated 90 degrees so that the audiovisual relationship was reversed. In Experiment 2, irrelevant sounds were interspersed between the speech sounds. (Color figure online)

horizontally and vertically in alternation. The ellipses consisted of a black, unfilled outline (0.34 cd/m^2; 0.21° thick) surrounded inside and out by red (CIE[.417, .318]) or green (CIE[.327, .418]) borders (15.5 cd/m^2; 0.10° thick). Each ellipse was treated with a Gaussian blur (0.12° radius), as this was found to facilitate spatially extended interocular suppression (see Fig. 1a; upper panel).

Depending on the eye of presentation, each ellipse would elongate either horizontally or vertically over a 265-ms period, remain static for 250 ms at full extension (2.11° × 1.17°), and then transition smoothly to the alternate orientation over another 265-ms period, where they would remain static for another 250 ms. This animation was repeated 35 times for each trial, resulting in 36 second trials. Throughout each trial, the ellipses presented to each eye followed the same dynamics but were shifted 180 degrees out of phase from each other. Thus, the image presented to one eye was always a 90-degree rotation of the image presented to the other eye, with a differently colored border. Importantly, because the sizes of the stimuli were always matched, the total contrast energy of the stimuli presented to each eye was always balanced. To aid binocular fusion, the stimuli were always surrounded by a binocularly presented high-contrast frame (3.72° × 3.72°; 0.74° thick) and presented on a gridded background made of overlapping gray bars. A black fixation cross (0.34 cd/m^2; 0.17° × 0.17°) was overlaid over the center of the stimuli presented to each eye.

Auditory stimuli consisted of two amplitude-balanced 200-ms recordings (presented at ~62 dB SPL) of a male speaker pronouncing the phonemes /wi:/ and /wo /. For each cycle of the 1,030 ms ellipse animation (see Fig. 1b), the /wi:/ sound was presented after 190 ms and the /wo / sound was presented after 700 ms. This timing was chosen to approximate the natural temporal correspondences between the articulatory motions and speech sounds. Exact timing was chosen by varying auditory timing in two-frame (23.5 ms) intervals around our initial estimates, and selecting the timing that produced the strongest crossmodal biasing effect in authors J.P. and E.G.

To facilitate recovery from adaptation to the experimental stimuli, participants were presented between trials with a 1-minute sample of video and sound from a film of ocean waves crashing on a rocky shore. The video was presented binocularly (8.67° × 4.87°) through the same stereoscope used in the experimental trials. This video was selected because it contained a wide variety of motion signals at different spatial frequencies and strong auditory spectral dynamics. Further, this video served as a foil for the purpose of the experiment, with many participants reporting in postexperimental interviews that they suspected this video to be part of the critical manipulation in our experiment.

Procedure Participants first completed six 36-second practice trials without sound to ensure that they understood the instructions and could experience rivalry. For the first practice trial, participants viewed a simulated rivalry display in which a binocularly presented circle alternated in color between red and green. Participants were instructed to continuously report the perceived color of the circle throughout the trial by holding down a corresponding key on a keypad, but not to press anything if they experienced a piecemeal or ambiguous percept, or if the two colors appeared equally dominant. The experimenter ensured that participants were responding correctly and gave additional instruction if they were not. In the second practice trial, static red and green circles were presented separately to each eye, triggering binocular rivalry. Participants were told that their perception might vary more unpredictably during this and subsequent trials, but to continue following the previous instructions as consistently and accurately as possible. Finally, participants were presented with the animated ellipses used in the experimental trials, but without the corresponding sounds, in four practice trials (counterbalancing color and ellipse dynamics across eyes). Participants were instructed to ignore the deformations of the ellipses and to focus only on reporting the color of the ellipse that they perceived to be dominant at any given time. Specifically, they were instructed to report an ellipse as dominant if it appeared to either occlude or suppress the other ellipse. After the practice trials, participants were shown the 1-minute video of ocean waves and instructed to view the entire video every time it was presented.

Participants were then presented with sixteen 36-second experimental trials including both the animated ellipses and the speech sounds. The eye and color in which congruent and incongruent stimuli appeared were counterbalanced and randomized across trials. Participants were instructed to ignore the sounds as distracters and to focus only on the visual dominance task as described in the practice trials.

Following the experiment, participants completed an open-ended interview designed to assess their knowledge of the audiovisual congruency present in the experimental stimuli. The interview was centered on three primary questions. First, they were asked whether the visual stimuli ever reminded them of or appeared to them as anything other than simple ellipses. Second, they were asked whether they felt that the auditory stimuli influenced their responses at all. Third, they were asked if they noticed any relationship between the auditory and visual stimuli. Participants were encouraged to respond with as much detail as possible.

Results

To assess whether perceptual competition in binocular rivalry was biased toward audiovisually congruent percepts, we compared the average proportion of each trial duration that participants reported perceiving the congruent and incongruent percept as dominant (cumulative dominance proportions). Results are shown in Fig. 2. Participants reported perceiving the congruent ellipse (M = 48.2% of total trial duration) as dominant for a significantly greater proportion of the trial duration than the incongruent ellipse (M = 44.3%), $t(11) = 2.53$, $p < .05$, $d = 0.73$, suggesting that visual awareness was biased toward the audiovisually congruent percept.

None of the participants' responses in postexperimental interviews suggested that they recognized the relationship between the visual and auditory stimuli. First, no participants reported that the visual stimuli reminded them of speaking mouths or anything else with a potential relationship to the auditory stimuli. Rather, participants reported interpreting the stimuli as other auditorily irrelevant percepts, such as blooming flowers. Second, participants did not report that the auditory stimuli influenced their responses at all, except for temporary distraction from the binocular rivalry task. Finally, no participants reported recognizing the relationship between the auditory and visual stimuli. Some participants reported noticing the synchrony between the dynamics of the ellipses and the timing of the sounds. However, recognition of this temporal relationship would not be sufficient to alert participants to the phonetic congruency exhibited by only one of the stimuli. Because the two animations expanded and contracted at the same rate, any temporal synchronicities applied equally to both visual stimuli.

Collectively, these results suggest that visual awareness can be biased toward audiovisually congruent percepts, even when participants do not recognize the congruency. To provide stronger support for this claim, we conducted a second experiment that incorporated a more stringent behavioral test of congruency recognition.

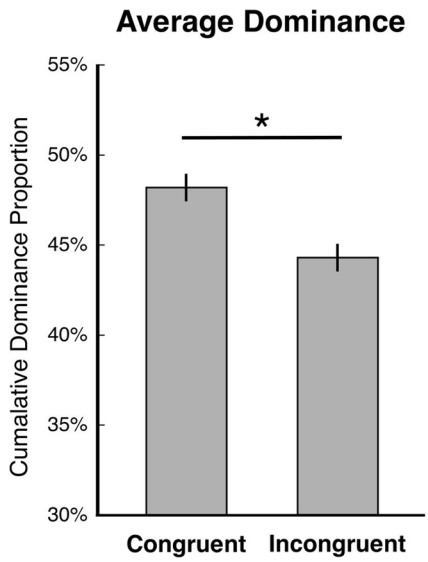

Fig. 2 Group average cumulative dominance proportions for each percept during binocular rivalry in Experiment 1. On average, the audiovisually congruent stimuli were dominant for a larger proportion of each trial than the incongruent stimuli. *Error bars* represent ±1 *SEM* adjusted for the repeated-measures design of the experiment. *$p < .05$

Experiment 2

Our second experiment differed from our first experiment in three important ways. First, in our first experiment, we used artificial visual stimuli that were only later matched with recorded speech. To improve ecological validity and potentially strengthen the observed effect, we derived the movements of the ellipses in Experiment 2 from the actual articulatory dynamics produced by the speaker of the auditory speech stimuli. Because this method made audiovisual congruencies more apparent, we also presented distractor sounds between the speech sounds to reduce the impression that there was any relationship between the auditory and visual stimuli.

Second, we introduced a control condition in which speech sounds were presented asynchronously with both visual stimuli. This allowed us to ensure that any observed effects were due to audiovisual congruency and not to differences in the visual stimuli or to the presentation of auditory signals in general. Note that, because attentional diversion toward even unrelated sounds can influence the dynamics of binocular rivalry (Alais, van Boxtel, Parker, & van Ee, 2010), a control condition that includes sounds that are virtually identical to those in the experimental condition (except for temporal misalignment) is preferable to one without sounds.

Finally, we used a more stringent behavioral test of participant's recognition of audiovisual congruency, and included only participants who did not perform significantly better than chance on this task in our analysis.

Method

Participants Fourteen undergraduate and graduate students at Northwestern University gave informed consent to participate in exchange for $15. They all had normal or corrected-to-normal visual acuity and normal color vision. Each was tested individually in a dimly lit room.

Apparatus Visual and auditory stimuli were presented using the same equipment in the same light-dimmed room as in Experiment 1.

Stimuli Visual and auditory stimuli were both derived from simultaneous audiovisual recordings of the same male speaker. To ensure smooth, continuous transitions in mouth shape, the speaker slowly articulated each syllable over a period of 560 ms, with 280-ms unvoiced periods between phonemes. Visual cues were used to facilitate proper timing, and the best timed 1,680-ms recording was extracted and used as the basis for the experimental stimuli.

Visual stimuli were created by fitting ellipses to the lips of the speaker in each frame of the selected video clip. Ellipses were subsequently scaled and centered so that they could be presented in a manner similar to that used in Experiment 1. They were then tinted and filled with red or green (with the same luminance and chromaticity as in Experiment 1) and treated with a Gaussian blur (0.06° radius) to facilitate spatially extended suppression. Because stimuli were wider at maximal horizontal extension (1.71° × 0.71°) than they were tall at maximal vertical extension (0.84° × 1.44°), we created the incongruent ellipses by rotating congruent ellipses by 90 degrees. This allowed us to ensure that the total contrast energy presented to each eye was matched at all times. The stimuli were always surrounded by a binocularly presented noise-patterned frame of the same size as the frame used in Experiment 1 and were presented on a dark background (10.1 cd/m^2) with black gridlines (see Fig. 1a). A black fixation cross was overlaid over the center of the stimuli presented to each eye.

Auditory speech stimuli were modified by replacing the silent periods between voiced phonemes with randomly ordered irrelevant sounds (e.g., automobile traffic, bubbling sounds, rain). These sounds were used to add plausibility to experimental instructions, which stressed that the presented sounds were mere distractors and were to be ignored. To minimize amplitude modulations in the sounds (presented at ~62 dB SPL), all speech and distractor sounds were amplitude normalized, with the first and last 20 ms of each sound crossfaded linearly.

For each trial, the 1,680-ms visual and auditory stimuli were presented continuously for 28 cycles, producing 47-s trials. In audiovisually synchronous trials, the sounds were presented such that the speech sounds were congruent with the animation presented to one eye, but not the other. An irrelevant sound was played before the first speech sound and between subsequent speech sounds. In the audiovisually asynchronous trials, the sounds were shifted 280 ms out of sync with both animations by removing the irrelevant sound from the beginning of the sequence. This temporal offset was selected because it placed the auditory speech stimuli outside of the typical window of audiovisual integration for speech (van Wassenhove, Grant, & Poeppel, 2007). Thus, the asynchronous condition contained the same visual stimuli (the unrotated and rotated ellipse animations), but neither visual stimulus was naturally synchronized with the speech sounds. In the postexperimental congruency detection task (described below), participants were presented with two continuous cycles (3,360 ms) of both the synchronous and asynchronous audiovisual stimuli used in the binocular rivalry experiment.

Procedure Participants were given the same instructions and completed the same practice trials (modified to include the new visual stimuli) as in Experiment 1.

Participants completed 20 trials with synchronous and 20 trials with asynchronous sounds in a random order. The color and eye to which audiovisually congruent (unrotated) and

incongruent (rotated) ellipses were presented were counterbalanced and randomized across each trial type. As in Experiment 1, participants viewed and listened to a 1-minute video of waves crashing on a rocky shore after each trial.

After completing the binocular rivalry experiment, participants performed a two-interval two-alternative-forced-choice (2I2AFC) task designed to assess whether or not they could detect the audiovisual congruency present in the synchronous trials. On each of 32 trials, participants were presented sequentially with two shortened (3,360 ms; two cycles of the animation) clips of the audiovisual stimuli used in the binocular rivalry task. One clip was always from the synchronous condition, whereas the other was from the asynchronous condition (order randomized). The eye and color in which congruent (unrotated) and incongruent (rotated) stimuli appeared were counterbalanced and randomized across trials. Participants were instructed to identify the interval containing the audiovisual congruency by keypress on each trial. We reasoned that if participants recognized the speech-based relationship between the audiovisual stimuli, then they would be able to complete this task with ease. However, if they did not recognize the audiovisual relationship, they would likely perform at chance levels.

The experimenter explained to participants that, during one interval, some feature of the soundtrack would correspond to one of the ellipses and not the other, but that during the other interval, that same feature would not correspond to either ellipse. Participants were instructed to try to identify the critical auditory feature and to report the interval that contained the crossmodal correspondence on each trial. The experimenter explained that on any given trial, the critical auditory feature could correspond with either the red or the green ellipse, and that the participant's task was to identify the interval in which the auditory feature corresponded with either ellipse, regardless of color. Last, the experimenter verified that the participant understood the instructions and answered any clarifying questions that did not provide information about the crossmodal correspondence of interest.

Results

To ensure that any effects observed in binocular rivalry could not be due to participants' recognition of audiovisual congruities, we first analyzed participants' performance on the postexperimental congruency detection task. Participants whose performance surpassed the significance criterion $\alpha = .05$ in a one-tailed binomial test (i.e., 21 or more correct responses out of 32 trials, falling in the upper 5% of the binomial distribution, with $n = 32$ and chance = .5) were considered to be likely to have detected the audiovisual congruency and were thus removed from further analysis.

Note that this is a liberal criterion for recognition of audiovisual congruency during the binocular rivalry task because, in the detection task (a) we alerted participants to the existence of the congruency and encouraged them to find it, whereas they were instructed to ignore sounds during the binocular rivalry task; (b) participants did not have to allocate attention to the perceptual dominance task, allowing them to devote all of their attention to congruency detection; and (c) explicit recognition of the critical shape–sound congruency would likely lead to near-perfect performance, well above our rejection criterion. Thus, participants who were not rejected based on this criterion were very likely to not have detected the audiovisual congruency during the binocular rivalry task.

Six participants performed significantly above chance (50%) on the congruency detection task, with a mean performance of 87.0% ($SD = 4.4\%$). The remaining eight participants had a mean accuracy of 42.6% ($SD = 5.8\%$), suggesting that they could not detect the audiovisual congruencies present in our experiment. We therefore continued our analysis of the binocular rivalry task using only the data from those eight participants.

Results are shown in Fig. 3a. On trials with audiovisually synchronous stimuli, participants reported perceiving the congruent ellipse ($M = 47.5\%$) as dominant for a significantly greater portion of the trial duration than the incongruent ellipse ($M = 42.2\%$), $t(7) = 3.36$, $p < .05$, $d = 1.19$. These results suggest that visual awareness was biased toward the audiovisually congruent percept, even when participants did not recognize the congruency. In contrast, on trials with audiovisually asynchronous stimuli, participants were no more likely to perceive either the unrotated ($M = 43.5\%$) or rotated ($M = 44.3\%$) ellipse as dominant, $t(7) = 0.22$, $p = .83$. These results suggest that the results observed in the synchronous condition were unlikely to have been driven by unaccounted for differences in the visual stimuli or by the presentation of auditory signals in general.

To illustrate the consistency of this result, dominance differences for asynchronous trials (proportion unrotated dominant – proportion rotated dominant; vertical axis) are plotted against the corresponding dominance differences for synchronous trials (proportion congruent dominant – proportion unrotated dominant; horizontal axis) in Fig. 3b. These difference scores provide an index of the relative dominance of the two competing stimuli in each condition, with a score of zero indicating equal dominance. As evidenced by the clear rightward clustering of the individual data points, synchronous auditory stimulation produced a consistent perceptual bias toward the audiovisually congruent stimulus. In contrast, as evidenced by the absence of upward or downward clustering of the data points, asynchronous auditory stimulation produced no consistent perceptual bias toward the unrotated or rotated stimulus.

To assess the temporal characteristics of the congruency effect in more detail, we separately compared the frequency

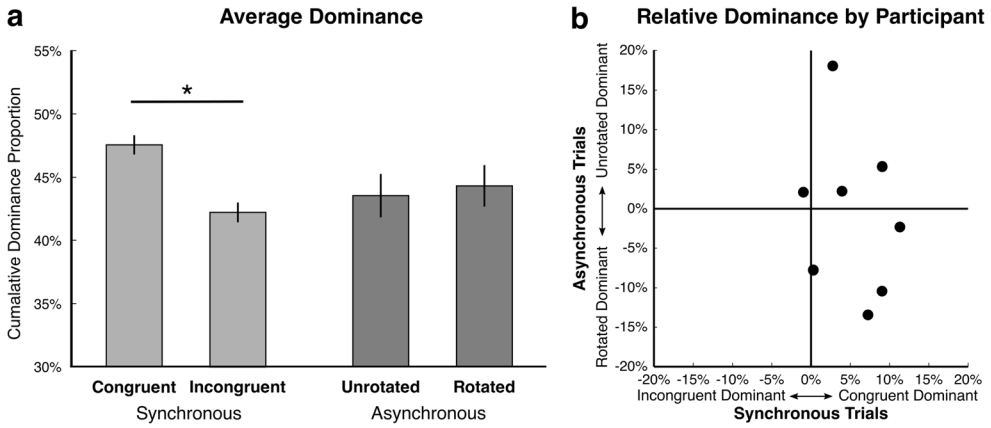

Fig. 3 Results from Experiment 2. **a** Group average cumulative dominance proportions during binocular rivalry. When the audiovisual timing matched that of natural speech on synchronous trials, the congruent (and unrotated) stimulus was dominant for a larger proportion of each trial than the incongruent (and rotated) stimulus (*lighter bars on the left*). However, when the auditory stimuli were presented 280 ms early on asynchronous trials, neither the unrotated or rotated stimulus was statistically more dominant (*darker bars on the right*). *Error bars* represent ±1 *SEM* adjusted for the repeated-measures design of the experiment (*$p < .05$). **b** Relative dominance of the two competing percepts in the synchronous (*horizontal axis*) and asynchronous (*vertical axis*) conditions for individual participants. Each point represents a single participant, with the horizontal and vertical lines intersecting the origin representing equal dominance in each condition. Increased rightward displacement indicates greater dominance of the congruent stimulus compared to the incongruent stimulus on synchronous trials, while increased upward displacement indicates greater dominance of the unrotated stimulus compared to the rotated stimulus on asynchronous trials. While rightward clustering is evident, upward clustering is absent

and average duration of individual dominance periods for each percept. On synchronous trials, neither the congruent (M = 9.79 dominance periods per trial) nor the incongruent stimulus (M = 9.67) was perceived as dominant more frequently, $t(7) = 0.63$, $p = .55$. However, when the congruent ellipse was dominant, it remained so for significantly longer (M = 2.72 s) than the incongruent ellipse did when it was dominant (M = 2.48), $t(7) = 2.93$, $p < .05$, $d = 1.03$. Because cumulative dominance duration is the product of percept frequency and dominance duration, these results suggest that the observed crossmodal effects on cumulative dominance proportion are the result of prolonged dominance periods for the congruent percept. By contrast, there were no significant differences in percept frequency ($M_{unrotated}$ = 9.84, $M_{rotated}$ = 9.79), $t(7) = 0.69$, $p = .51$, or duration ($M_{unrotated}$ = 2.59, $M_{rotated}$ = 2.59), $t(7) < 0.01$, $p > .99$, on asynchronous trials.

Last, as an additional test to evaluate whether recognition of crossmodal congruency contributed to the observed crossmodal disambiguation effect, we assessed whether participants' scores on the congruency detection task predicted the size of their crossmodal congruency effect during binocular rivalry. If congruency detection played a role in the auditory effect on binocular rivalry, then participants with higher congruency detection scores (proportion correct detection of synchronous vs. asynchronous stimuli) would be expected to show larger congruency effects during binocular rivalry (proportion congruent dominant – proportion incongruent dominant on synchronous trials). However, congruency detection scores were, if anything, negatively correlated with the size of participants' congruency effects during binocular rivalry; $r(12) = -0.45$, $p = .11$ for all 14 participants, and $r(6) = -.47$, $p = .24$, for the eight participants included in the main analysis of binocular rivalry. Thus, there was no evidence that recognition of auditory–visual congruence played a role in generating the auditory–visual congruency effect.

General discussion

We showed in two separate experiments that sounds biased visual competition in binocular rivalry toward audiovisually congruent percepts, even when these congruencies were not apparent to participants. These results suggest that sounds can automatically disambiguate visual perception by facilitating awareness of audiovisually consistent percepts. These effects were observed in participants who did not identify the critical congruency in an open-ended interview (Experiment 1) or detect it in a simple 2AFC task (Experiment 2), suggesting that the observed effect was not the result of voluntary selective attention or response bias toward audiovisually congruent stimuli.

These results provide strong support for the growing literature suggesting that information integration across sensory modalities can resolve ambiguities within a particular sensory domain (Klink et al., 2012; Schwartz et al., 2012). While previous research suggests that tactile (Lunghi et al., 2010), olfactory (Zhou et al., 2010), and auditory information can resolve visual ambiguities, the current study added rigorous control for potential biases introduced by participant's knowledge of multisensory congruencies. By ruling out these potential attentional or response biases, our results provide evidence

for direct and automatic crossmodal disambiguation of the contents of visual awareness.

While our results clarify previous results by demonstrating that genuine crossmodal disambiguation can occur even when these confounding factors are controlled for, our experiments most likely provide only a conservative estimate of the potential strength of this type of effect. Covertly influencing participants' perception without alerting them to the crossmodal correspondences present in our stimulus display necessitated the use of artificial stimuli that lacked many of the multisensory cues typically exploited in the perception of naturalistic audiovisual speech (e.g., Jiang, Alwan, Keating, Auer, & Bernstein, 2002; Yehia, Rubin, & Vatikiotis-Bateson, 1998). While a stimulus display including a wider variety of audiovisually correspondent cues would potentially produce a stronger disambiguation effect, it would also increase the likelihood that participants would notice the crossmodal correspondence, potentially biasing their attention or response patterns. Thus, ensuring that these potential biases were rigorously controlled for limited our use of stimuli to those expected to produce statistically reliable but potentially modest congruency effects. Nevertheless, our results demonstrate that, when all other relevant factors are balanced, sounds reliably facilitate the perception of congruent visual information at the expense of ambiguous or alternative percepts, thereby disambiguating the contents of visual awareness. Further research is needed to ascertain the extent to which genuine crossmodal disambiguation, voluntary selective attention, and interactions between the two contribute to the sometimes larger congruency effects reported in previous work (e.g., Conrad et al., 2013).

Recent debates in the literature on multisensory perception have centered on the question of whether audiovisual integration in speech perception occurs automatically or relies on other factors, such as attention (Navarra, Alsius, Soto-Faraco, & Spence, 2010), awareness (Plass, Guzman-Martinez, Ortega, Grabowecky, & Suzuki, 2014), or recognition of speech stimuli as speech related (Tuomainen, Andersen, Tiippana, & Sams, 2005; Vroomen & Stekelenburg, 2011). While visual modulation of auditory speech perception is one of the most studied topics in the field of multisensory perception, far less research has focused on the converse—auditory effects on visual speech perception (Alsius & Munhall, 2013; Baart & Vroomen, 2010; Palmer & Ramsey, 2012; Sweeny et al., 2012). Our results corroborate and extend results suggesting that auditory speech can facilitate visual perception of congruent visual speech, even when participants are unaware of audiovisual congruencies (Alsius & Munhall, 2013; Palmer & Ramsey, 2012).

Because the rivaling stimuli in our experiments were matched for all features (dynamic and static) except for orientation and color, they would have been matched for general audiovisual congruencies not specific to speech, such as potential relationships between auditory amplitude envelope or fundamental frequency and visual luminance, size, shape, motion onset, or velocity. Thus, the effect demonstrated here was likely due to the speech-related congruency between auditory spectral dynamics and the directionality of shape deformations. If so, these results suggest that speech-related shapes and sounds in particular can be automatically integrated without an explicit recognition of their relationship. Future research is needed to identify the specific crossmodal correspondences that the perceptual system exploits to produce such effects and to see whether these correspondences also contribute to visual facilitations of auditory speech perception (e.g., Grant & Seitz, 2000; Sumby & Pollack, 1954).

References

Alais, D., van Boxtel, J. J., Parker, A., & van Ee, R. (2010). Attending to auditory signals slows visual alternations in binocular rivalry. *Vision Research, 50*(10), 929–935. doi:10.1016/j.visres.2010.03.010

Alsius, A., & Munhall, K. G. (2013). Detection of audiovisual speech correspondences without visual awareness. *Psychological Science, 24*(4), 423–431. doi:10.1177/0956797612457378

Baart, M., & Vroomen, J. (2010). Do you see what you are hearing? Cross-modal effects of speech sounds on lipreading. *Neuroscience Letters, 471*(2), 100–103. doi:10.1016/j.neulet.2010.01.019

Baier, R., Idsardi, W. J., & Lidz, J. (2007). *Two-month-olds are sensitive to lip rounding in dynamic and static speech events.* Presented at the Audiovisual Speech Processing Conference, Hilvarenbeek, Netherlands. Retrieved from http://ling.umd.edu/labs/acquisition/papers/BaierIdsardiLidz-AVSP07.pdf

Blake, R., & Logothetis, N. K. (2002). Visual competition. *Nature Reviews Neuroscience, 3*(1), 13–21. doi:10.1038/nrn701

Blake, R. R., Fox, R., & McIntyre, C. (1971). Stochastic properties of stabilized-image binocular rivalry alternations. *Journal of Experimental Psychology, 88*(3), 327.

Chen, Y.-C., Yeh, S.-L., & Spence, C. (2011). Crossmodal constraints on human perceptual awareness: Auditory semantic modulation of binocular rivalry. *Frontiers in Psychology, 2*. doi:10.3389/fpsyg.2011.00212

Chong, S. C., Tadin, D., & Blake, R. (2005). Endogenous attention prolongs dominance durations in binocular rivalry. *Journal of Vision, 5*(11), 6. doi:10.1167/5.11.6

Conrad, V., Bartels, A., Kleiner, M., & Noppeney, U. (2010). Audiovisual interactions in binocular rivalry. *Journal of Vision, 10*(10), 27. doi:10.1167/10.10.27

Conrad, V., Kleiner, M., Bartels, A., Hartcher, O.'. B. J., Bülthoff, H. H., & Noppeney, U. (2013). Naturalistic stimulus structure determines the integration of audiovisual looming signals in binocular rivalry. *PLoS ONE, 8*(8), e70710. doi:10.1371/journal.pone.0070710

Deroy, O., Chen, Y.-C., & Spence, C. (2014). Multisensory constraints on awareness. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 369*(1641), 20130207. doi:10.1098/rstb.2013.0207

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences, 8*(4), 162–169. doi:10.1016/j.tics.2004.02.002

Grant, K. W., & Seitz, P.-F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America, 108*(3), 1197–1208. doi:10.1121/1.1288668

Guzman-Martinez, E., Ortega, L., Grabowecky, M., Mossbridge, J., & Suzuki, S. (2012). Interactive coding of visual spatial frequency and auditory amplitude-modulation rate. *Current Biology, 22*(5), 383–388. doi:10.1016/j.cub.2012.01.004

Jiang, J., Alwan, A., Keating, P. A., Auer, E. T., & Bernstein, L. E. (2002). On the relationship between face movements, tongue movements, and speech acoustics. *EURASIP Journal on Advances in Signal Processing, 2002*(11), 506945. doi:10.1155/S1110865702206046

Kang, M.-S., & Blake, R. (2005). Perceptual synergy between seeing and hearing revealed during binocular rivalry. *Psichologia, 32,* 7–15.

Kim, Y. J., Grabowecky, M., & Suzuki, S. (2006). Stochastic resonance in binocular rivalry. *Vision Research, 46*(3), 392–406.

Klink, P. C., van Wezel, R. J. A., & van Ee, R. (2012). United we sense, divided we fail: Context-driven perception of ambiguous visual stimuli. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 367*(1591), 932–941. doi:10.1098/rstb.2011.0358

Lack, L. C. (1974). Selective attention and the control of binocular rivalry. *Perception & Psychophysics, 15*(1), 193–200. doi:10.3758/BF03205846

Lunghi, C., Binda, P., & Morrone, M. C. (2010). Touch disambiguates rivalrous perception at early stages of visual analysis. *Current Biology, 20*(4), R143–R144. doi:10.1016/j.cub.2009.12.015

Navarra, J., Alsius, A., Soto-Faraco, S., & Spence, C. (2010). Assessing the role of attention in the audiovisual integration of speech. *Information Fusion, 11*(1), 4–11. doi:10.1016/j.inffus.2009.04.001

Navarra, J., Yeung, H. H., Werker, J. F., & Soto-Faraco, S. (2012). Multisensory interactions in speech perception. In B. E. Stein (Ed.), *The new handbook of multisensory processing* (pp. 435–452). Cambridge: MIT Press.

Palmer, T. D., & Ramsey, A. K. (2012). The function of consciousness in multisensory integration. *Cognition, 125*(3), 353–364. doi:10.1016/j.cognition.2012.08.003

Parker, A. L., & Alais, D. M. (2006). Auditory modulation of binocular rivalry. *Journal of Vision, 6*(6), 855–855. doi:10.1167/6.6.855

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science, 6*(2), 191–196. doi:10.1111/1467-7687.00271

Plass, J., Guzman-Martinez, E., Ortega, L., Grabowecky, M., & Suzuki, S. (2014). Lip reading without awareness. *Psychological Science, 25*(9), 1835–1837. doi:10.1177/0956797614542132

Schwartz, J.-L., Grimault, N., Hupé, J.-M., Moore, B. C. J., & Pressnitzer, D. (2012). Multistability in perception: Binding sensory modalities, an overview. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 367*(1591), 896–905. doi:10.1098/rstb.2011.0254

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*(2), 212–215. doi:10.1121/1.1907309

Sweeny, T. D., Guzman-Martinez, E., Ortega, L., Grabowecky, M., & Suzuki, S. (2012). Sounds exaggerate visual shape. *Cognition, 124*(2), 194–200. doi:10.1016/j.cognition.2012.04.009

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio–visual speech perception is special. *Cognition, 96*(1), B13–B22. doi:10.1016/j.cognition.2004.10.004

van Ee, R., van Boxtel, J. J. A., Parker, A. L., & Alais, D. (2009). Multisensory congruency as a mechanism for attentional control over perceptual selection. *The Journal of Neuroscience, 29*(37), 11641–11649. doi:10.1523/JNEUROSCI.0873-09.2009

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia, 45*(3), 598–607. doi:10.1016/j.neuropsychologia.2006.01.001

Vatakis, A., Ghazanfar, A. A., & Spence, C. (2008). Facilitation of multisensory integration by the "unity effect" reveals that speech is special. *Journal of Vision, 8*(9), 14. doi:10.1167/8.9.14

Vatakis, A., & Spence, C. (2007). Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli. *Perception & Psychophysics, 69*(5), 744–756. doi:10.3758/BF03193776

Vatakis, A., & Spence, C. (2008). Evaluating the influence of the "unity assumption" on the temporal perception of realistic audiovisual stimuli. *Acta Psychologica, 127*(1), 12–23. doi:10.1016/j.actpsy.2006.12.002

Vroomen, J., & Stekelenburg, J. J. (2011). Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition, 118*(1), 75–83. doi:10.1016/j.cognition.2010.10.002

Wheatstone, C. (1838). On some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical Transactions of the Royal Society of London, 128,* 371–394.

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication, 26*(1/2), 23–43. doi:10.1016/S0167-6393(98)00048-X

Zhou, W., Jiang, Y., He, S., & Chen, D. (2010). Olfaction modulates visual perception in binocular rivalry. *Current Biology, 20*(15), 1356–1358. doi:10.1016/j.cub.2010.05.059